ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 115−122

Journal of Clinical Epidemiology

VARIANCE AND DISSENT

Synthesis of observational studies should consider credibility ceilings

Georgia Salanti[a], John P.A. Ioannidis[a,b,*]

[a]*The Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina, School of Medicine, Ioannina, Greece*

[b]*The Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts University, School of Medicine, Boston, MA, USA*

Accepted 27 May 2008

Abstract

Objective: Meta-analyses of observational studies often get spuriously precise results. We aimed to factor this skepticism in meta-analysis calculations.

Study Design and Setting: We developed a simple sensitivity analysis starting from the assumption that any single observational study cannot give us more than a maximum certainty *c%* (called *credibility ceiling*) that an effect is in a particular direction and not in the other. Each study included in meta-analysis is adjusted for different credibility ceilings *c* and the consistency of the conclusion examined. We applied the method in three meta-analyses of observational studies with nominally statistically significant summary effects (mortality with teaching versus nonteaching health care; risk of non-Hodgkin's lymphoma with hair dyes; mortality with omega-3 fatty acids).

Results: Between-study heterogeneity I^2 estimates dropped from 36%−72% without a ceiling effect to 0% with ceilings of 9%, 4%, and 4% in the three meta-analyses, respectively. Nominal statistical significance was lost with ceilings of 10%, 8%, and 11%, respectively. The likelihood ratios suggested that even with minimal ceiling effects, there was no strong support for the credibility of each of these three associations.

Conclusions: Consideration of credibility ceilings allows conservative interpretation of observational evidence and can be applied routinely to meta-analyses of observational studies.

Keywords: Meta-analysis; Likelihood; Credibility; Precision; Bias

1. Introduction

Meta-analysis methods have been extensively used in the synthesis of data from observational studies. Despite the fact that findings from observational studies are often subject to bias [1], they are still the best available evidence in many research fields. By synthesizing a large amount of data from several observational studies, meta-analyses may occasionally reach summary effects with very tight confidence intervals (CIs). However, this precision is spurious given the inability to accommodate the extensive sources of confounding and other biases that lurk, often unmeasured and unobserved, in nonrandomized designs [2]. Accommodating for between-study heterogeneity (e.g., with random effects models) may inflate the CIs of the summary effects, but may still fail to account for all the uncertainty that accompanies the estimates of epidemiological studies.

One way to address this problem of spurious precision is to evaluate the impact of assuming different levels of bias affecting either the point estimate of each study or the variance thereof [3]. Several formal approaches have been proposed, mostly with a Bayesian framework [4]. One of the earliest proposed methods has been the confidence profile method advocated by Eddy et al [5,6]. Different biases, such as misclassification or measurement error, can be explicitly modeled as extra parameters in the likelihood, and their uncertainty may also be taken into account [7]. The method was initially developed for adjusting and combining pieces of evidence that come from randomized trials; applications regarding different study designs (including epidemiological studies) have followed [8]. However, explicit modeling of bias requires assumptions to be made regarding the nature of bias to introduce the appropriate parameters (e.g., modeling the rates of misclassified cases as controls). Moreover, familiarity with the Bayesian framework is required to apply the confidence profile method.

In this article, we apply a simple adjustment of the likelihood of the included studies which can be routinely performed in any frequentist software and without making

* Corresponding author. Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. Tel.: +302-651-097807; fax: +302-651-097867.

E-mail address: jioannid@cc.uoi.gr (J.P.A. Ioannidis).

0895-4356/09/$ − see front matter
doi: 10.1016/j.jclinepi.2008.05.014

explicit assumptions about bias. The basic (and only) assumption is that, given the methodological limitations implicit in its design, an observational study, no matter how large and well conducted, cannot give us more than a maximum certainty that an effect is in a particular direction and not null or in the other direction. This means that the likelihood of an effect being in a specific direction cannot exceed a specific level. This level, here called *credibility ceiling*, reflects a property of the study design and the topic of interest. The credibility ceiling is taken into account in modifying accordingly the likelihood function of the data from each study. One then examines a range of ceiling values for all studies or for different subgroups of studies that are considered to have different ceilings due to different designs. We perform a sensitivity analysis based on different degrees of credibility; for each ceiling value(s), the modified likelihood functions are synthesized in meta-analyses and we estimate the between-study heterogeneity and the summary effects. Then, one can plot these results as a function of the credibility ceiling values and interpret the results according to the ceiling values that are considered plausible.

2. Methods

2.1. Background

Consider the synthesis of n studies. We assume that the outcome of each study i is normally distributed and summarized by an effect size y_i with variance v_i. This is a reasonable assumption for large studies as is the typical case with epidemiological cohorts and also in the specific examples that we consider in this article. Then, assuming a random effects meta-analysis model, the study-specific likelihood $L_i(\mu, \theta_i, \tau^2|y_i, v_i)$ is a function of the underlying random effect θ_i, the common mean effect μ, and the heterogeneity variance τ^2. Then, the likelihood function of the meta-analysis model is $L(\mu, \boldsymbol{\theta}, \tau^2|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^{n} L_i(\mu, \theta_i, \tau^2|y_i, v_i)$.

By DerSimonian and Laird, it follows that $\hat{\mu} = \sum y_i w_i / \sum w_i, \quad w_i = 1/(v_i + \hat{\tau}^2)$ [9]. Although random effects analysis ''inflates'' the variance of the pooled effect $\hat{\mu}$ to take into account heterogeneity, that is, $\mathrm{var}(\hat{\mu}) = 1/\sum 1/(v_i + \hat{\tau}^2)$, the resulting CIs may still be insufficiently narrow to account for the methodological caveats in the individual studies.

2.2. Incorporating a credibility ceiling

A major assumption underlies our method: in a single epidemiological study, there is at least c probability that the underlying effect is not in the direction of the effect suggested by the observed point estimate y_i. It is assumed that this credibility ceiling of c cannot be reduced further, regardless of how large and meticulous the study is, because there are unsurpassable limitations inherent in its design and the topic under study that preclude higher levels of certainty. Thus, it is assumed that a single study of this type can never give more than $(1-c)/c$ certainty that the effect is in the direction suggested by the point estimate versus not in this direction, if an effect does exist. If this level of certainty must be exceeded, several studies replicating an effect in that direction need to be seen repeatedly. To account for this, the random variable $u_i \sim N(y_i, v_i)$ is considered. Then, the credibility $P(u < 0|y_i > 0)$ or $P(u > 0|y_i < 0)$, that is, the probability that the variable takes values on the opposite direction of the observed point estimate, is calculated. If this probability is less than c, the variance v_i^* is recalculated as $v_i^* = \max\{(y_i/z_c)^2, v_i\}$ with z being the inverse of the cumulative normal distribution. The new modified study specific likelihood functions $L_i(\mu, \theta_i, \tau^2|y_i, v_i^*)$ are then synthesized according to typical inverse variance meta-analysis procedures and the variance of the pooled effect is further inflated as $\mathrm{var}(\hat{\mu}) = 1/\sum 1/(v_i^* + \hat{\tau}^2), \quad v_i^* \geq v_i$. Note that the heterogeneity parameter would also be altered as it depends on the study-specific variances.

2.3. Sensitivity analysis

Determination of the credibility ceiling c is rather arbitrary and can be based on prior beliefs or expert opinion. Using a range of plausible values is suggested to observe the changes in the summary estimate and heterogeneity. This sensitivity analysis will yield useful information, such as the estimation of the maximum credibility ceiling beyond which the current conclusion of the meta-analysis would be challenged. Note that the credibility ceiling c may be the same for all studies incorporated in the meta-analysis, or different ceilings may be used; studies with different designs may imply different levels of the maximum certainty that can be achieved and, therefore, different ceiling values.

2.4. Summary of the method

The method can be summarized in the following steps:

1. In each study, assuming that there is an effect equal to the observed effect size, calculate the probability that an observed effect with sampling variance equal to v_i would be on the opposite direction of the true effect.
2. If this probability is less than a predefined credibility ceiling c%, inflate the variance $v_i = (y_i/z_c)^2$.
3. Do meta-analysis using the inflated variances.
4. Repeat steps 1–3 for a range of plausible credibility ceiling values.

2.5. Software

All calculations presented in this article have been performed in R [10]. A readily used module for applying different ceilings can be downloaded from the software page at www.dhe.med.uoi.gr.

3. Application

The methods in the analysis of 3 systematic reviews on 3 important topics that have been published recently in major journals are outlined. These meta-analyses have found a nominally statistically significant summary effect even with standard random effects calculations ($p < 0.05$) [11−13]. The 3 illustrative examples are representative of situations where the level of statistical significance in the original meta-analysis ranges from close to 0.05 to much lower (0.024−0.004), and they have a wide range in the number of synthesized studies (3−74).

3.1. Patient outcomes with teaching versus nonteaching health care

A large meta-analysis has synthesized data from observational studies examining whether teaching versus nonteaching health care services have differences in mortality [11]. The meta-analysis synthesized relative risks across 74 comparisons using an inverse variance random effects method. The summary effect was marginally statistically significant, with point risk ratio estimate of 0.96 and 95% CI of 0.93−1.00 ($p = 0.024$). There was large heterogeneity between studies ($I^2 = 72\%$, $p < 0.001$ for Cochran's Q statistic). The authors interpreted the nominally statistically significant effect as suggestive that there is no clear difference in mortality rates between teaching and nonteaching health care. The rationale was that these studies had a large uncertainty in their results that went beyond what could be measured and reflected in the CIs [8]. Mortality is influenced by a myriad of risk factors, most of which were not measured in any study. Moreover, most studies used administrative databases rather than medical records. These typically not only lacked information on even major confounders, but also had large room for measurement and information bias due to the rudimentary quality of the data. In particular, by their very nature (automated recording of many thousands of admissions), administrative databases can easily become large and the uncertainty in their effect estimates can be spuriously minimal. Studies based on administrative databases may thus have far more weight in the calculations than studies based on meticulous examination of full medical records−unless something is done to downweight the spurious precision.

Here, the meta-analysis is re-performed considering different values for the credibility ceiling c. In an illustrative analysis, let us use $c = 25\%$ for all studies. This is our best bet from a clinical standpoint on what is the minimal ceiling uncertainty that one should expect in these studies. It is considered that even if teaching and nonteaching health care do differ in mortality, it is impossible that a single study (no matter how well done and what its results are) can ascertain more than 75% that teaching health care has better outcomes than nonteaching health care or vice versa. This means that a single study can never make us believe that the superiority of one type of health care is $(1 - 0.25)/0.75 = 3$ times more likely than the superiority of the other type of health care. The random effects summary estimate under this assumption becomes 0.99 (95% CI, 0.97−1.01) and the point estimate of I^2 changes from 72%−0%.

Let us now consider a range of c values. A range of credibility ceilings were used from 1% to 40% with a step of 1%. As expected, the heterogeneity I^2, which depends on the study size [14], drops sharply (with $c = 1\%$, it drops from 72%−47%) and it reaches $I^2 = 0\%$ for a ceiling of 9% (Fig. 1a). The magnitude of the summary point estimate diminishes with increasing ceilings. The 95% CIs for the odds ratio become slightly narrower up to ceiling 10% because of the drop in the heterogeneity; thereafter they become wider. In all, for a ceiling of 10% or larger, the 95% CIs exclude a relative risk reduction (RR) exceeding 4% with the provision of health care in teaching rather than nonteaching facilities. The likelihood that the effect is in favor of teaching hospitals rather than in favor of nonteaching hospitals decreases steeply with increasing ceiling effect after $c = 10\%$ (Fig. 1c). It may also be reasonable to attribute different degrees of credibility to studies according to whether the information in the studies comes from clinical as contrasted to administrative data. The subgroup summary estimates in the original meta-analysis were 0.98 (95% CI, 0.95−1.00) for administrative data and 0.92 (95% CI, 0.84−1.02) for clinical data. Figures 2a and b present the change in heterogeneity and likelihood ratio for different combinations of ceilings. There are 49 studies with administrative data and 25 studies with clinical data; therefore, the drop in heterogeneity and increase in likelihood ratio are sharper for the same ceiling in the former. Very large ceilings are reasonable to assume for studies based on administrative databases, where the quality of the data is typically limited and the ability to adjust for important confounders is poor. Figure 2b shows the likelihood ratio for death in teaching health care compared with nonteaching health care. As shown, for most plausible combinations of ceiling values in clinical and administrative database studies, the likelihood ratio remains below the level of strong support (<32). A likelihood ratio exceeding 32 requires a zero ceiling for the clinical data, if the administrative data ceiling is 20%, and it is impossible to achieve if the ceiling of the administrative data is higher than 20%.

One may also split the data into more than 2 categories with different ceilings. To illustrate, the data were split into 4 categories of studies with different ceilings: 25% for administrative data with results that have not been adjusted for major confounders (volume/experience, severity, and comorbidity), 10% for administrative data with results adjusted for these major confounders or for clinical data that have not been adjusted, and 5% for adjusted clinical data. The pooled estimate was 0.98 (95% CI, 0.97−1.00) with estimated $I^2 = 0$; the likelihood ratio of teaching health care being better was 17.7. One can also calculate the

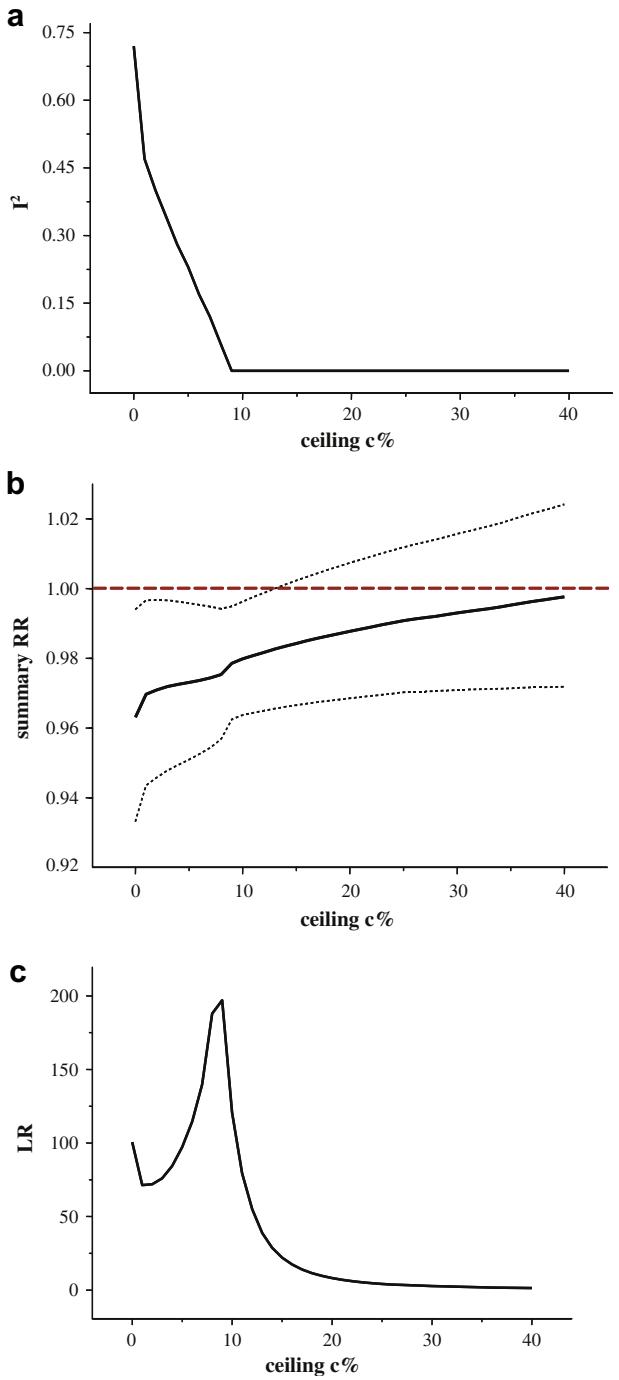

Fig. 1. Mortality with teaching versus nonteaching health care. In the horizontal axis are plotted a range of ceiling values (0%–40%). The vertical axis shows in the three plots, respectively, (a) the heterogeneity metric I^2 [14], (b) the summary risk ratio with 95% CIs (also shown is the line of null effect), and (c) the likelihood ratio of having better outcome in teaching health care over better outcome in nonteaching health care.

likelihood ratio for the effect being larger than a specific value that may be considered clinically worthwhile. There may be some unavoidable subjectivity on what exactly is clinically worthwhile, but, for example, with these values of ceilings, the likelihood ratio of teaching health care achieving at least a 4% relative RR in mortality was as low as 0.004.

3.2. Hair dyes and non-Hodgkin lymphoma

Another highly visible meta-analysis found some possible association between the personal use of hair dyes and hematopoietic cancers, with non-Hodgkin's lymphoma (NHL) showing the strongest association [12]. The summary relative risk for NHL was 1.23 (95% CI, 1.07–1.42, $p = 0.0043$) and there was an heterogeneity index ($I^2 = 55\%$). The 14 studies included in the meta-analysis attributing different ceilings were reanalyzed. The heterogeneity first reaches an estimate of 0% at a ceiling of 4% (Fig. 3a); nominal statistical significance is lost at a ceiling of 8% (Fig. 3b); and the relative likelihood of a harmful rather than beneficial effect drops from 461 in the original meta-analysis without ceiling effect to below 32 at a ceiling of 9% (Fig. 3c). The relative likelihood of a 20% relative risk increase for NHL (relative risk 1.2) drops below 0.03 at a ceiling of 1%.

3.3. Omega-3 fatty acids and mortality

A very interesting example of meta-analysis of observational studies whose conclusion has not been supported by the meta-analysis of randomized trials on the same topic has been published in [13]. Three cohort studies have shown a significant drop in mortality in the group receiving omega-3 fatty acids compared with the control (RR = 0.65 (95% CI, 0.48–0.88, $p = 0.006$), $I^2 = 36\%$), whereas 13 randomized trials have found no statistically significant benefit (RR = 0.87 [95% CI, 0.73–1.03]).

Application of different ceiling values to the cohort data shows that the estimate of I^2 becomes 0% with a very small ceiling of 4%. The summary risk ratio estimate loses its nominal statistical significance with a ceiling of 11% (Fig. 4b). The relative likelihood of the protective effect compared with the harmful decreases sharply from the original value of 314 (without any ceiling) and becomes less than 32 at a ceiling of 12% (Fig. 4c). The relative likelihood of a protective effect representing at least a 10% relative RR (RR = 0.9) is more than 32 only if the ceiling is 4% or less.

4. Discussion

Meta-analysis is widely used to synthesize the results of epidemiological studies despite the threat of the underlying biases. A simple approach is proposed that can be routinely applied to evaluate the consistency of the results when different degrees of maximum credibility are considered for the presence of an effect in the observed findings of single studies. The proposed approach may help in the more conservative interpretation of the results of epidemiological meta-analyses and avoid spurious precision and unjustified

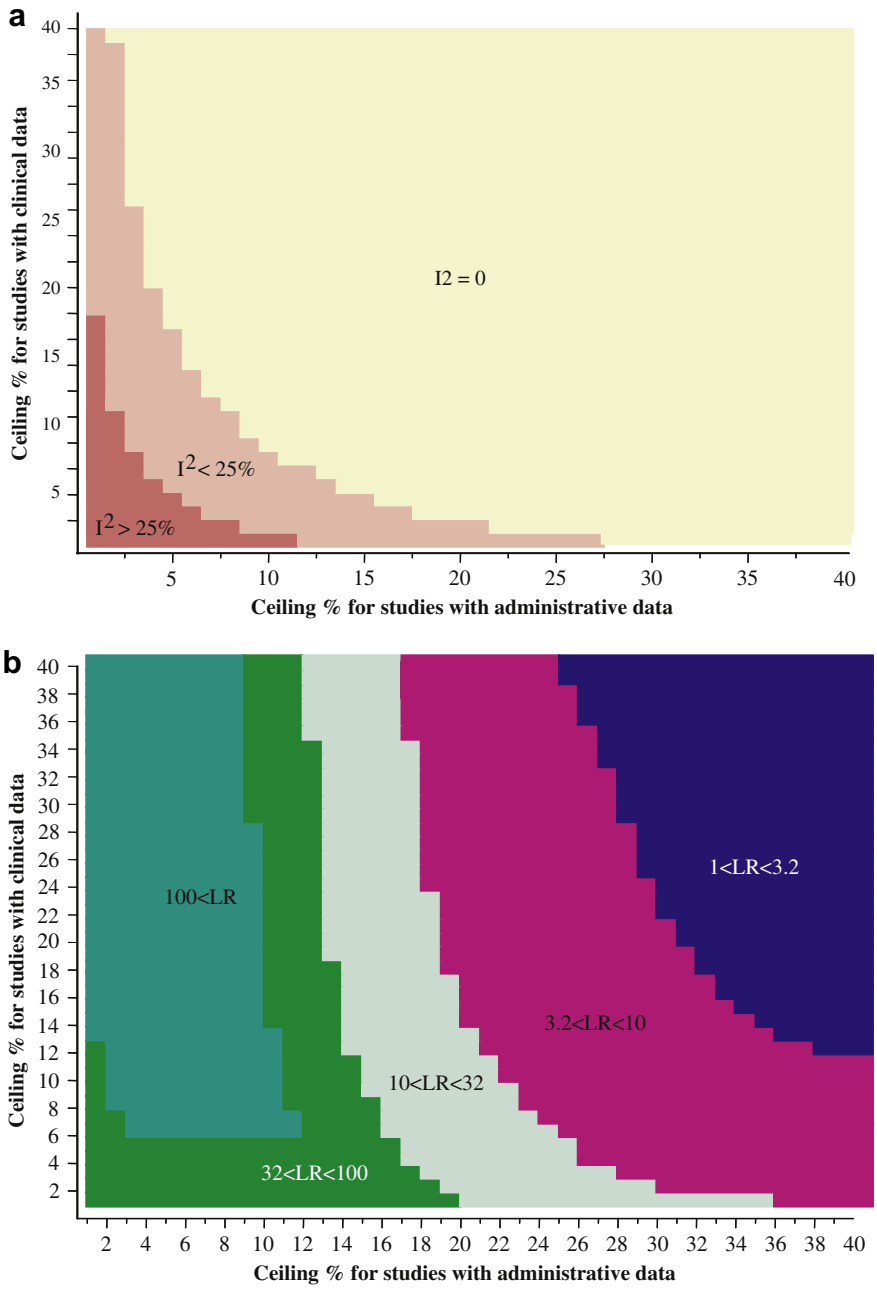

Fig. 2. Mortality with teaching versus nonteaching health care. In the horizontal axis are plotted a range of ceiling values (0%–40%). The vertical axis shows in the two plots, respectively, (a) the heterogeneity metric I^2 in groups of zero, low ($I^2 < 25\%$) and average to high heterogeneity ($I^2 > 25\%$), and (b) the likelihood ratio of having better outcome in teaching health care over better outcome in nonteaching health care for combinations of different credibility ceilings for studies with administrative data and studies with clinical data.

claims for significant effects based on nominal statistical significance.

Bias in the primary studies cannot be corrected with statistical manipulation, so our approach represents an alternative to limit the precision of estimates. The proposed credibility ceiling value is limited to the direction of the effect rather than skepticism as to the magnitude of the effect. The use of this approach in sensitivity analyses could be a way of showing readers and scientists how skeptical they should be given the results of a meta-analysis rather than just verbalizing its limitation. More than a quantitative solution to spurious findings from a meta-analysis of observation studies, this is suggested as a didactic tool that authors can use to ensure that readers take the findings with the necessary caution.

The 3 meta-analysis examples that have been analyzed show that nominal statistical significance may be misleading, when the constituent studies carry more uncertainty than what their CIs convey. Differences in mortality between teaching and nonteaching health care, harm from

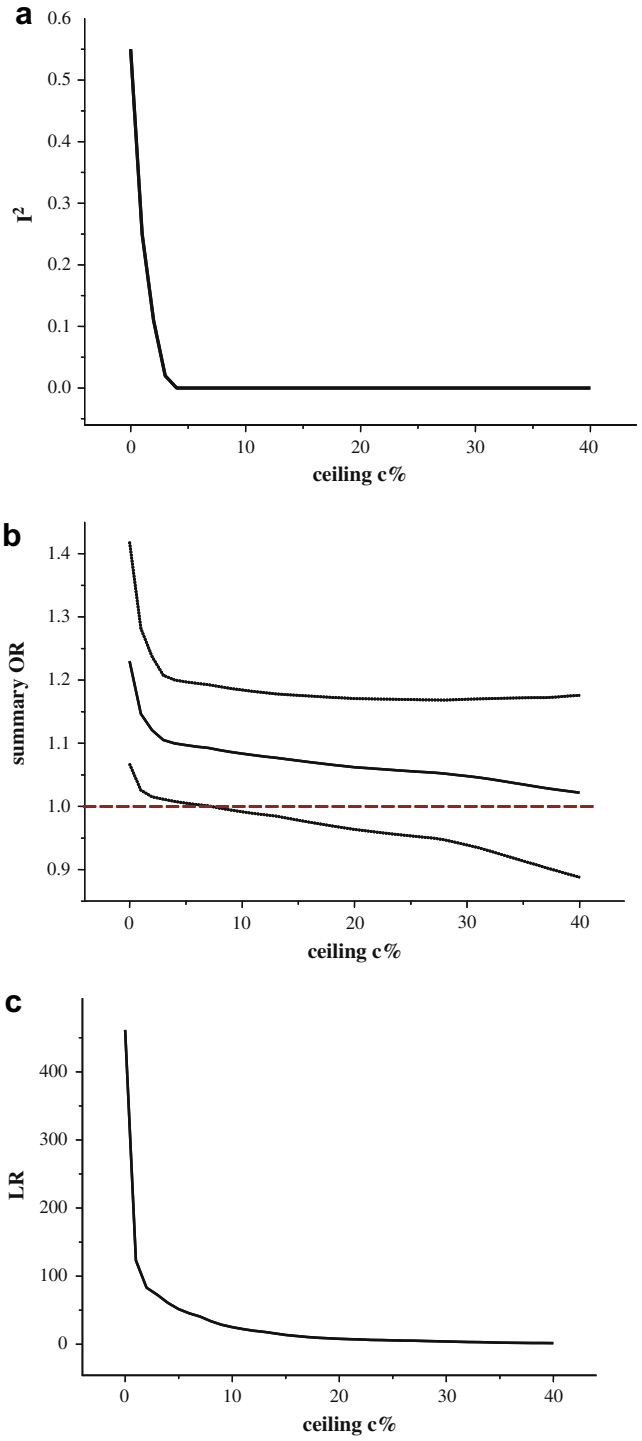

Fig. 3. Hair dyes and non-Hodgkin lymphoma. In the horizontal axis are plotted a range of ceiling values (0%–40%). The vertical axis shows in the three plots, respectively, (a) the heterogeneity metric I^2, (b) the summary risk ratio with 95% CIs (also shown is the line of null effect), and (c) the likelihood ratio of having higher risk with the use of dyes over higher risk among the non-users.

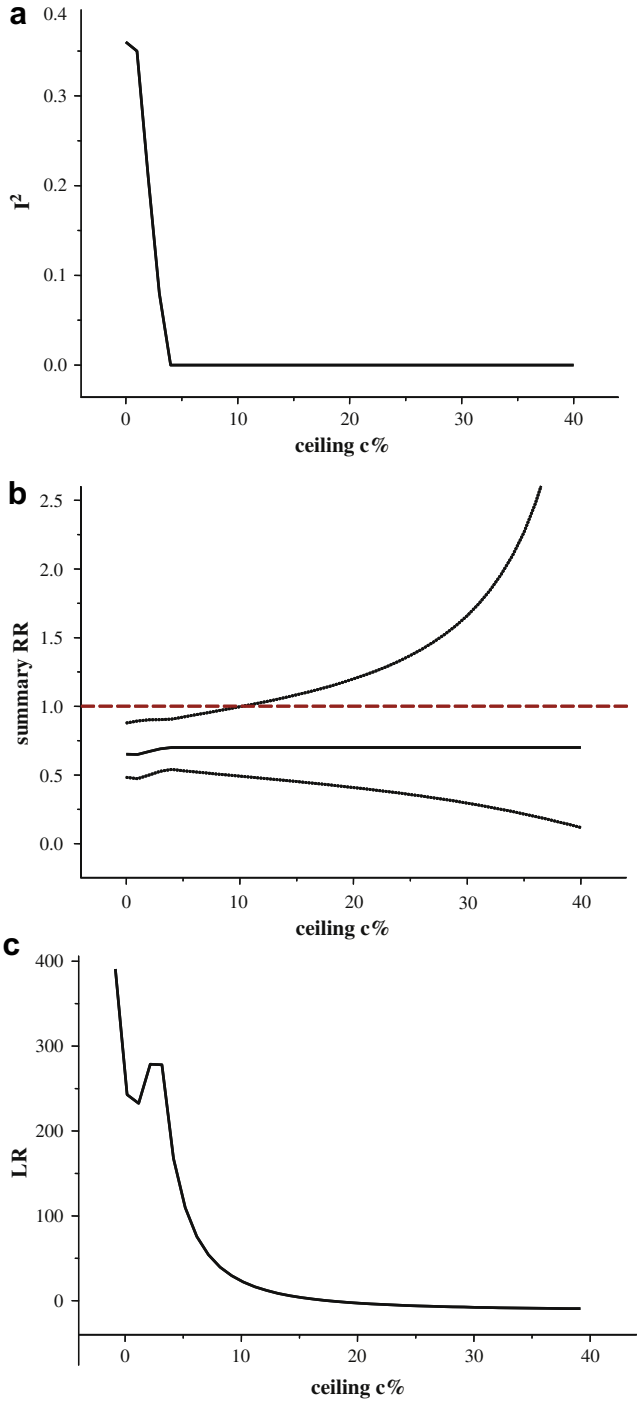

Fig. 4. Omega-3 fatty acids and mortality. In the horizontal axis are plotted a range of ceiling values (0%–40%). The vertical axis shows in the three plots, respectively, (a) the heterogeneity metric I^2, (b) the summary risk ratio with 95% CIs (also shown is the line of null effect), and (c) the likelihood ratio of having lower risk in the omega-3 group over lower risk in the control group.

hair dyes, and benefits from omega-3 fatty acids are probably not major, and they may not exist at all. Statistically significant findings are very common in published observational studies [2,15–17], and meta-analyses thereof also reach nominally statistically significant results [2,18]. In a setting where almost all single studies and meta-analyses being published find nominally statistically significant results, the criterion of nominal statistical significance probably loses much of its discriminating power to separate

true from spurious associations [19–21]. Thus, additional methods that probe the robustness of the observed associations must be used.

In all our analyses, the adjustments conferred by the consideration of different ceilings modify the spread of the likelihood distributions of single studies, but do not change the maximum likelihood (point) estimates of each study. Thus, the basic assumption is that the observed point estimates are not biased. Obviously, one may wish to challenge this assumption and also examine what would happen with modifications in the point estimates. However, many investigators might object to changing the maximum likelihood results of a study. Conversely, the tight CIs often are clearly overprecise, as in the case of large studies that use data of questionable accuracy and validity. As discussed in the introduction, several other methods exist for explicitly modifying both the point estimates and spread of the likelihood. They are more complex than the ceiling method proposed, but they may be worthwhile considering, when such modifications are justified.

Another caveat is that with the proposed approach, between 2 studies that have different observed effects, but the same variance, the variance is inflated more for the study that shows an effect that lies further away from the null. Thus, the proposed method tends to penalize studies that show more extreme effects than those that show effects closer to the null. Given this peculiarity, the summary effects are more likely to shrink toward null than increase with increasing ceiling values.

Finally, the ceiling considerations create a framework where replication of an association across many studies becomes very important. With large ceilings, studies tend to have relatively similar weights in the meta-analysis calculations. Reaching nominal statistical significance for an effect in ceiling-adjusted calculations requires the conduct of many studies showing consistent effects for an association. Again, this is reasonable because it has been argued that consistent replication is key for ascertainment of proposed epidemiological associations [22]. Given the wide diversity of populations and settings that can be encountered in most epidemiological studies, more extensive replication before adopting an association as being credible is probably warranted. However, a potential disadvantage of large ceilings is that among studies with similar observed effects, large studies are penalized more than smaller studies. This makes sense in cases such as the teaching health care meta-analysis in which large studies were those where the data were likely to have the worst quality. However, on other occasions the larger studies may have better quality than smaller studies. Unfortunately, determination and grading of the quality of single epidemiological studies is usually very difficult or even impossible, and most investigators argue with good reason against weighting studies differently based on quality weights [23].

The range of values of c that may be investigated is unavoidably a subjective choice. For example, one may argue that values of c exceeding 10% may be too skeptical. However, a considerable range of c values is worth investigating, and then, each reader and scientist can interpret the results based on this continuum.

Allowing for these caveats, adoption of ceilings offers a simple method that can be used routinely to probe and avoid the potentially spurious precision of meta-analyses of observational studies. We suggest that meta-analyses of observational studies factor this skepticism using the proposed sensitivity analysis and convey it regularly with expressions of the kind ‘‘when we consider that there is no chance that any single study can convince us more than c% that the effect of the exposure is harmful/beneficial, the pooled estimate becomes…’’ and so forth. This allows a more cautious interpretation of the results of meta-analyses of observational studies.

References

[1] Deeks JJ, Dinnes J, D’Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii–173.

[2] Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34.

[3] Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. BMJ 1996;313:603–7.

[4] Spiegelhalter DJ, Abrams KR, Myles PJ. Evidence synthesis. In: Spiegelhalter DJ, Abrams KR, Myles PJ, editors. Bayesian approaches to clinical trials and health-care evaluation. Chichester: John Wiley & Sons Ltd; 2004.

[5] Hasselblad V, Eddy DM, Kotchmar DJ. Synthesis of environmental evidence: nitrogen dioxide epidemiology studies. J Air Waste Manage Assoc 1992;42:662–71.

[6] Eddy DM, Hasselblad V, Shachter R. A Bayesian method for synthesizing evidence. The confidence profile method. Int J Technol Assess Health Care 1990;6:31–55.

[7] Critchfield GC, Eddy DM. A confidence profile analysis of the effectiveness of disulfiram in the treatment of chronic alcoholism. Med Care 1987;25(Suppl 12):S66–75.

[8] Wolpert R, Mengersen K. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. Stat Sci 2004;19:450–71.

[9] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.

[10] R [computer program]. 2006.

[11] Papanikolaou PN, Christidi GD, Ioannidis JP. Patient outcomes with teaching versus nonteaching healthcare: a systematic review. PLoS Med 2006;3:e341.

[12] Takkouche B, Etminan M, Montes-Martinez A. Personal use of hair dyes and risk of cancer: a meta-analysis. JAMA 2005;293:2516–25.

[13] Hooper L, Thompson RL, Harrison RA, Summerbell CD, Ness AR, Moore HJ, et al. Risks and benefits of omega 3 fats for mortality, cardiovascular disease, and cancer: systematic review. BMJ 2006;332:752–60.

[14] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.

[15] Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. PLoS Med 2007;4:e79.

[16] Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 2004;329:883.

[17] Kyzas P, Denaxa-Kyza D, Ioannidis J. Almost all cancer prognostic marker studies report statistically significant results. Eur J Cancer 2007;43:2559−79.

[18] Kyzas PA, Cunha IW, Ioannidis JP. Prognostic significance of vascular endothelial growth factor immunohistochemical expression in head and neck squamous cell carcinoma: a meta-analysis. Clin Cancer Res 2005;11:1434−40.

[19] Wacholder S, Chanock S, Garcia-Closas M, El GL, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96:434−42.

[20] Sterne JA, Davey SG. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226−31.

[21] Ioannidis JPA. Why most published research findings are false. PLoS Medicine 2005;2:e124.

[22] Moonesinghe R, Khoury MJ, Janssens AC. Most published research findings are false-but a little replication goes a long way. PLoS Med 2007;4:e28.

[23] Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001;2:463−71.